DOCUMENT RESUME

ED 028 147 SP 002 446

By-Gage, N.L.; And Others

Explorations of the Teacher's Effectiveness in Explaining. Technical Report No. 4, Stanford Center for Research and Development in Teaching.

Stanford Univ., Calif. School of Education.
Spons Agency-Office of Education (DHEW), Washington, D.C. Cooperative Research Program.
Bureau No-BR-5-0252
Pub Date Dec 68
Contract-OEC-6-10-078
Note-59p.; Chapter in Research into Classroom Processes, ed. Ian Westbury; pub. The Ontario Institute for Studies in Education (in press).
EDRS Price MF-$0.50 HC-$3.05
Descriptors-Behavior Rating Scales, *Communication Skills, Computer Oriented Programs, Content Analysis, Educational Experiments, *Effective Teaching, Evaluation Criteria, Lesson Observation Criteria, Media Research, Statistical Analysis, Teacher Behavior, *Teacher Evaluation
Identifiers-Attention Test, Stanford Teacher Competence Appraisal Guide

This document presents four correlated studies based on (1) the concept of "micro-criteria" which narrows the dimensions of investigating teacher effectiveness through the variable (explaining), the potential correlates (classroom behavior), and the rating of effectiveness (pupil achievement) and (2) data from an initial experiment in which 12th grade teachers taught two 15-minute lessons to their pupils who subsequently took a comprehension test and rated teachers and themselves on performance with an adapted Stanford Teacher Competence Appraisal Guide and Attention Test. Study 1 emphasizes the statistical methods used in determining the reliability, generality, and correlation (all found to be positive) of teacher effectiveness and performance. Study 2 investigates which type of lesson recording would yield ratings closest to actual classroom ratings and which teacher behaviors pupils observed with a free-response instrument. The latter investigation's finding, that good teaching and cognitive activities are consistently related, is supplemented in study 3 which categorizes 27 teachers behaviors (such as rule-example-rule presentation) in an attempt to apply objective measurement to rating teacher effectiveness . The final study discusses the use of computer programs to improve reliability in boring or complex tasks such as word counting. (LP)

ED028147

STANFORD CENTER
FOR RESEARCH AND DEVELOPMENT
IN TEACHING

Technical Report No. 4

EXPLORATIONS OF THE TEACHER'S EFFECTIVENESS IN EXPLAINING

N. L. Gage, Maria Belgard, Daryl Dell,
Jack E. Hiller, Barak Rosenshine, and
W. R. Unruh

School of Education
Stanford University
Stanford, California

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

December 1968

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education, under the provisions of the Cooperative Research Program (Contract No. OE-6-10-078, Project No. 0102).

Authors of specific studies of the Explaining Project are identified within the body of this report. The studies were made independently on the same data, and no collective responsibility is implied.

SP002446

C O N T E N T S

EXPLORATIONS OF THE TEACHER'S EFFECTIVENESS IN EXPLAINING

EXPLORATIONS OF THE TEACHER'S EFFECTIVENESS IN EXPLAINING[1]

THE MICROCRITERION OF EFFECTIVENESS IN EXPLAINING

N. L. Gage, Stanford Center for
Research and Development in Teaching

The four studies described in this report explored the role of the teacher as explainer. The studies were based on a modification of the "criterion of effectiveness" paradigm. This paradigm, used successfully in other areas of research, has been dominant in the study of teaching. The paradigm is one in which the investigator undertakes to:

> Identify or select a criterion (or a set of criteria) of teacher effectiveness. This criterion then becomes the dependent variable. The research task is then (1) to measure this criterion, (2) to measure potential correlates of this criterion, and (3) to determine the actual correlations between the criterion and its potential correlates (Gage, 1963, p. 114).

In many studies, the criterion has been a rating of "teaching effectiveness" assigned to the teacher by the school principal, and the potential correlates have been measures of the personality and characteristics of the teacher. Many attempts to identify good teachers have been made in this manner; hundreds of studies yielding thousands of correlation coefficients have been carried out. But reviewers of this research have not been impressed. "In the large, these studies have yielded disappointing results: correlations that are non-significant, inconsistent from one study to the next, and usually lacking in psychological and educational meaning" (Gage, 1963, p. 118). Similarly, Getzels and Jackson (1963), after reviewing research on teacher personality, concluded that:

> Despite the critical importance of the problem and a half-century of prodigious research effort, very little is known for certain about the nature and measurement of teacher personality and teaching effectiveness. The regrettable fact is that many of the studies so far have not produced significant results (p. 574).

---

[1]Adaptations of these papers were presented at a symposium at the meetings of the American Educational Research Association in February 1968. This report will be included as a chapter in Research into Classroom Processes, edited by Ian Westbury and Arno A. Bellack, to be published by The Ontario Institute for Studies in Education and the Teachers College Press, Columbia University.

Under such circumstances, researchers have sought to refine the paradigm, shifting to different types of potential correlates and criteria. In place of personality traits, the potential correlates have become objectively denoted classroom behaviors of pupils and teachers; in place of a global rating of effectiveness, the criterion has become a measure of pupil achievement. Researchers who followed this approach (e.g., Medley and Mitzel, 1959; Spaulding, 1963; Soar, 1966) have usually conducted their research during a school year, sampling the teacher's behavior two to five times during the year, and otherwise imposing no restrictions on materials to be used or objectives to be achieved.

"Micro-Criteria" of Effectiveness

A further refinement in the criterion-of-effectiveness paradigm was suggested by Gage (1963). He advocated reducing the complexity of the problem through the use of "micro-criteria" of effectiveness:

> Rather than seek criteria for the overall effectiveness of teachers in the many, varied facets of their roles, we may have better success with criteria of effectiveness in small, specifically defined aspects of the role. Many scientific problems have eventually been solved by being analyzed into smaller problems whose variables were less complex (Gage, 1963, p. 120).

The studies conducted by Flanders (1965) and Bellack (1966) may be regarded as exemplifying an approach toward micro-criteria. Both investigators reduced their criteria to pupil achievement over a short period (two weeks and one week, respectively) and attempted to standardize the investigation by providing all teachers with identical and previously unused curricular materials.

But it may be argued that their criteria were still highly complex. Many aspects of the teacher's role were being considered; the designs permitted a wide range of behaviors. That is, in the investigations of Flanders and Bellack, the teacher was expected to (a) motivate the reading of prepared materials, (b) promote discussion, and (c) maintain discipline as well as (d) engender a cognitive grasp of the material. The teachers were also allowed unrestricted freedom to choose their own methods, and the consequent differences in methods made comparisons difficult. Some teachers spent most of the class time lecturing, while others promoted decision-making by the students.

It is too early to draw conclusions as to the most effective design for determining the correlates of teaching effectiveness because too few studies have been completed. But the concept of micro-criteria of effectiveness suggests that we study teaching by using criteria even simpler than those used by Flanders and Bellack. Simplifications might include the reduction of teaching time, focus upon only one type of teaching behavior and its outcomes, and further control of teaching procedures.

The Teacher's Effectiveness in Explaining

The present four studies are attempts to investigate teacher effectiveness with the "micro-criterion" of effectiveness approach. The specific teacher behavior selected for investigation was explaining. "Explaining" is the skill of engendering comprehension--usually orally, verbally, and extemporaneously--of some process, concept, or generalization. Explaining occurs in all grade levels and subject matters, whether it is a fifth-grade teacher explaining why the time in New York differs from that in San Francisco or a geologist explaining how the ice age may have been caused by volcanic eruptions. Everyday observation tells us that some people explain aptly, getting to the heart of the matter with just the right terminology, examples, and organization of ideas. Other explainers, on the contrary, get us and themselves all mixed up, use terms beyond our level of comprehension, draw inept analogies, and even employ concepts and principles that cannot be understood without an understanding of the very thing being explained. Explaining may come close to being the essence of instruction, so that when a teacher is attempting to explain proportionality to his geometry class or irony to his English class, he is behaving more purely as a teacher than when he is attempting, say, to motivate, promote discussion, or maintain discipline.

The teacher's "effectiveness in explaining" was defined operationally as the ability to present ideas in such a way that the pupils would be able to respond to questions testing the comprehension of those ideas. Explanation as defined by philosophers of science was not the focus of these studies. Rather, these investigations were concerned with the kind of pedagogical "explaining" discussed by Swift (1961), Thyne (1963, pp. 126-155), Meux and Smith (1964, pp. 146-148), Nuthall and Lawrence (1965, pp. 33-48), and Bellack and his associates (1966, pp. 24-25).

The same basic data were used in all four studies. After the initial data were gathered, the investigators studied specific problems. Study I, by Belgard, Rosenshine, and Gage, focussed on three questions: How reliable and general is the teacher's ability to explain? How reliable and general are pupil ratings? Which pupil ratings are related to the teacher's effectiveness in explaining?

Study II, by Unruh, dealt with the question, Which type of protocol, or record of the teacher's behavior, is the most effective source of cues that observers can use in rating teacher effectiveness? This study also included an exploration of the validity of free-response and structured ratings in an attempt to determine new variables, perceptible by pupils, which are related to the teacher's effects.

Study III, by Rosenshine, and Study IV, by Dell and Hiller, were investigations of specific behaviors whose frequency of occurrence might be related to teacher effectiveness in explaining. Both investigations were innovative in that they dealt with verbal behaviors not customarily measured in systematic studies of teaching behavior. In the study by Rosenshine, human coders counted the frequencies of the various syntactic, linguistic, and gestural events in teachers' behavior. In the study by Dell and Hiller, computers and specially developed dictionaries were used in the search for correlates of effectiveness; the capabilities of an IBM 7090 computer as a high-speed clerk were exploited. Different variables were measured in the two studies.

All of the studies were exploratory and correlational, and it would be hazardous to infer that the significant findings reflect causal relationships. Much experimental and correlational work will be necessary to confirm and expand the present findings. Nonetheless, the studies demonstrate a set of techniques that may yield knowledge on how to improve the effectiveness of classroom explanations.

Method

The basic records of teacher behavior and effects, which were used in all four studies, were collected by Rosenshine and Belgard. Fifty-eight experienced social studies teachers and their twelfth-grade classes in public schools in the San Francisco Bay Area participated in the study as volunteers. Only classes in which pupils were grouped heterogeneously according to general

ability were used; class size ranged from 10 to 31 with a mean size of 21. Complete data were available initially for 43 teachers and 898 pupils, but due to the deterioration or loss of some videotape recordings, the number of videotaped lectures on each topic ranged in the four studies from 26 to 38. There was nothing to suggest that the elimination of the recordings was not random.

All teachers taught lessons based upon identical materials. The materials were "Atlantic Reports" in issues of the Atlantic magazine between November 1964 and August 1965; they were judged by curriculum experts to be suitable for twelfth-grade social studies classes. The teachers were asked to explain the material in the reports, which dealt with economic, political, and social conditions in Yugoslavia and Thailand. The term "explain" was operationally defined as the process whereby a teacher's 15-minute lecture on the prescribed curriculum material would enable his students to answer 10 multiple-choice questions on the content. The pupils' mean adjusted score on the test was used as the index of the teacher's effectiveness in explaining. The adjustment procedures are explained below.

Standardization of Lesson Procedure

A week before his explanation of the material to his class, each teacher received a copy of (a) the Atlantic Reports on Yugoslavia and Thailand, (b) the five odd-numbered items of each multiple-choice test, (c) a modified version of the Stanford Teacher Competence Appraisal Guide, (d) the Attention Report, and (e) the following instructions:

> Explain the important ideas and principles contained in the article. You may organize the lecture in any way you wish. The 10 test items will probe comprehension of main ideas and not test for knowledge of little pieces of information. For your guidance, you will be given five of the ten questions which will be used on each of the student tests.
>
> Limit yourself to the content of the article. Do not do any additional reading or research on the topic covered by either article, or add material to your lecture. This rule is intended to save you unnecessary effort, and to insure that all teachers work with identical curricular material, no more and no less. The 10-item test will deal only with material in the article.
>
> Limit yourself to lecture and use of the chalkboard for purposes of this study. In subsequent studies, we may investi-

gate the effects of such other procedures as questioning, discussion, questions from students, study sheets, student note-taking, and use of various types of projectors. But in this study, if teachers use techniques other than lecture with the chalkboard, it will be impossible to perform an adequate analysis of the data. For some teachers this restriction may require a difficult departure from their customary teaching style. We hope that you will bear with us, since the purpose of this study is to investigate explaining behavior per se and not to evaluate any teacher or groups of teachers. We also hope, therefore, that you will discourage student questions during the lecture.

The 15-minute restriction is necessary to equate time conditions for all of the 50 teachers participating in this study. You may pace yourself, or you may have the equipment operator give you a signal when there are two or five minutes remaining. (The lecture must end after 15 minutes.)

During the 15-minute lecture period you will be in complete charge of the classroom. The Stanford research assistant will be available to aid you. You will signal the start of the lecture.

On the second day, the procedure will be exactly the same, except that you will lecture on the Thailand article.

On the third day, the procedure will vary slightly. To make possible the control of differences in student ability, all students will hear an identical 15-minute tape-recording of an article on Israel. The only responsibility of the teacher during the playing of this tape will be that of maintaining class order. The testing will be carried out as usual.

The Yugoslavia and Thailand lessons were taught by the teachers to their own pupils in their regular classrooms on each of the first two days of the period of the study. Videotape recordings were made of all the lessons. On the third day, an audiotape record of an Atlantic Report on Israel was played for all the classes. Immediately after each of the three lessons, a comprehension test, the Stanford Teacher Competence Appraisal Guide, and the Attention Report were administered by the investigators.

Measures

The comprehension tests consisted of 10-item multiple-choice tests constructed for a previous study (Fortune, Gage, and Shutes, 1966). The adjusted mean score of each class on each test was used as an index of the teacher's effectiveness in explaining the Atlantic Reports on Yugoslavia and Thailand.

Because there were no teacher effects in the audiotape-recorded lesson, presented to all classes, in Study I (by Belgard, Rosenshine, and Gage) the mean test scores on Israel were used to adjust the mean class scores on Yugoslavia and Thailand for differences between classes in aptitude for achievement on lessons of this type, and the ratings given to the Israel taperecorded-lecture were used to adjust the ratings given to the other lectures for tendencies to rate teachers favorably or to report high or low attention.

In Studies II, III, and IV, the raw mean posttest class scores on Yugoslavia and Thailand were adjusted for two predictor variables: (1) the mean test scores on Israel, and (2) a score developed by using content analysis procedures to assess the relevance of the material in each lecture. This second adjustment was made to insure that the differences between classes were not due merely to differences in the pertinence of the presentations of the teachers. In making all adjustments, the between-groups regression slope was used.

After each lesson, the pupils rated the lesson on an adaptation of the Stanford Teacher Competence Appraisal Guide, which deals with the following dimensions: (1) clarity of aims, (2) organization of the lecture, (3) beginning the lecture, (4) clarity of presentation, (5) pacing the lecture, (6) pupil attention, (7) ending the lecture, (8) teacher-pupil rapport, and (9) amount of learning. For each dimension, the ratings were made on a seven-point scale ranging from "truly exceptional" to "weak," with an additional category for "unable to observe." The class mean of the pupils' ratings for each of the nine dimensions was computed for each lesson, and this mean was used as an index of how the class rated the lesson on each dimension.

After each lesson, pupils also filled out the Attention Report, which solicited self-report information on four items, of which the following is an example:

> During this lecture, my mind wandered and I began to think about other things:
>
> 0. all of the time.
> 1. most of the time.
> 2. some of the time.
> 3. a little bit of the time.
> 4. none of the time.

The mean of the pupils' total scores on the four items was used as the attention score of the class for a particular lesson.

I. THE TEACHER'S EFFECTIVENESS IN EXPLAINING: EVIDENCE ON ITS GENERALITY AND CORRELATION WITH PUPILS' RATINGS AND ATTENTION SCORES[2]

Maria Belgard, Stanford Center for
Research and Development in Teaching

Barak Rosenshine, Temple University

N. L. Gage, Stanford Center for
Research and Development in Teaching

The teacher's effectiveness in explaining was defined in this study as the teacher's ability to present ideas to his pupils in such a way that the pupils would be able to respond to questions that assessed the comprehension of ideas. Given this definition, one can ask, How consistent is the teacher's effectiveness? Effectiveness in explaining in any given instance may be only a function of the particulars of that instance and hence predict nothing about effectiveness in other situations. But if effectiveness is general in some significant degree, generalizable findings as to its determiners, correlates, and consequences are more likely.

Results

The discussion of the results of this study will deal first with the reliability and generality of effectiveness in explaining over lessons; second, with the reliability and generality of pupils' mean ratings and attention reports over lessons; and third, with the correlations of effectiveness in explaining with class ratings of teacher competence and with measures of pupil attention.

Correlations Among Mean Comprehension Test Scores

The means, standard deviations, Horst coefficients, and intercorrelations (zero-order, partial, and part correlations) of the mean comprehension scores of the classes participating in this study are shown in Table 1. The Horst coefficients (rs = .77, .76, and .76) show that the three sets of mean scores differ from class to class with substantial reliability.

The three correlations among unadjusted mean scores indicate that classes with high mean scores on one test tended to get high mean scores on the other two. For example, the correlation between unadjusted mean scores on Yugoslavia

---

[2] The authors are indebted to Katherine Baker for editorial assistance.

Table 1

Intercorrelations among Total Mean Comprehension Scores, Unadjusted and Adjusted, on Yugoslavia and Thailand, and Israel
(N = 43 Classes Containing 10-31 Students)

Mean Score	Mean of Means	S. D. of Means	Horst Coefficient	(2) Thailand	(3) Israel	(4) Yugoslavia	(5) Thailand
1. Unadjusted Yugoslavia Total	6.91	.92	.77	.63[b]	.58[b]	.81[d]	.37[d]
2. Unadjusted Thailand Total	6.24	.76	.76		.52[b]	.40[d]	.84[d]
3. Unadjusted Israel Total	5.17	.81	.76			-.01[d]	-.03[d]
4. Adjusted Yugoslavia Total[a]							.47[c]
5. Adjusted Thailand Total[a]							

[a] Adjusted for Mean Score of Class on the Israel Comprehension Test

[b] Zero-order Correlation

[c] Partial Correlation

[d] Part Correlation

and unadjusted mean scores on Thailand was .63, which indicated that the proportion of variance among the means on one test predictable from the means on the other test is .40.

The predictable variance of the mean scores on Yugoslavia and Thailand can be seen as having two components, one which is specific to the Yugoslavia and Thailand scores and another which is shared by all three--the Yugoslavia, Thailand, and Israel scores. The variance which is common to all three scores is presumably due in large part to between-class differences in mean student ability, since no variance attributable to teachers entered into the tape-recorded lecture on Israel.

The correlation between the adjusted mean comprehension scores on Yugoslavia and Thailand (r = .47) represents the relationship between the mean scores on Yugoslavia and Thailand which remains when the effects of student ability and other irrelevant factors measured by the comprehension scores for Israel have been removed. Of the total proportion of variance common to both tests (namely, .40), the proportion which is common to all three lessons is .18. The proportion common only to the Yugoslavia and Thailand lessons, presumed to be due to teacher effects, at least in large part, is .22.

The part correlations shown in Table 1 indicate the relationship between the adjusted and unadjusted means. The part correlations involving the same variable are high (rs = .81 and .84) because the effects common to the live lessons and the lesson on Israel are not high. The part correlations between mean scores on Israel and adjusted mean scores on Thailand and Yugoslavia are, as expected, about zero since all the variance in Yugoslavia and Thailand scores which is predictable from Israel scores had been removed.

The other part correlations, which are less meaningful, are those of unadjusted mean scores on Yugoslavia with adjusted mean scores on Thailand (r = .37), and unadjusted mean scores on Thailand with adjusted mean scores on Yugoslavia (r = .40).

The foregoing method of estimating the generality of the teacher's effectiveness, as measured by the adjusted mean comprehension score, may not completely eliminate covariance due to the pupils' ability, since the same pupils are involved in both measures of mean comprehension. That is, the adjustment on the basis of the Israel mean may not completely eliminate covariance

irrelevant to a measure of the teacher's effectiveness. To obtain an estimate of generality less influenced by such irrelevant covariance, we made the following additional analysis. Each teacher's students were divided at random into odd-numbered and even-numbered students. Separate means were computed for the odd-numbered and even-numbered sub-classes on the Yugoslavia, Thailand, and Israel tests. The means on the Yugoslavia and Thailand tests were adjusted so as to eliminate the effects of variance between classes on the Israel test. Then correlations were computed between the adjusted mean of the odd-numbered students on Yugoslavia and that of the even-numbered students on Thailand (r = .16), and vice-versa (r = .38). (The full array of obtained rs is shown in Table 2.)

These two coefficients each represent the "reliability" of the adjusted mean score for a class half as large as those actually tested. To estimate from these coefficients the reliability of the adjusted means based on the entire class, the Spearman-Brown formula was applied. The adjusted rs obtained in this way were .28 and .55. The mean of the latter two rs is .41, which represents an estimate of the generality of teacher effectiveness over two different topics and two subsets of students whose differences in ability have been adjusted for. This estimate (r = .41) is lower than that obtained when the same students were involved (r = .47) because irrelevant covariance due to students was more effectively eliminated. The degree of generality reflected in this coefficient may be considered to characterize the teacher's effectiveness in explaining when both subject matter and pupils are varied.

Three results indicate that the teacher's effectiveness in explaining had some consistency across different topics and different groups of pupils. The correlations were not high enough to indicate that the effectiveness of individual teachers can be measured with adequate reliability with only two lessons, 10-item tests, and classes of about 21 students. For such reliability, higher than about .40, additional lessons, longer tests, and larger classes would be needed.

Correlations Among Pupils' Mean Ratings and Attention Scores

The second question concerned the reliability and consistency of pupil ratings. We were interested in knowing the degree to which the ratings were reliable, consistent from occasion to occasion, and influenced by general tendencies of pupils to give high or low ratings.

Table 2

Correlations between Unadjusted and Adjusted Mean Scores of Odd-Numbered and Even-Numbered Students on the Yugoslavia, Thailand, and Israel Tests (N = 43)

	Mean of Means	S. D. of Means	Unadjusted Mean Scores						Adjusted Mean Scores			
			Odd-Numbered			Even-Numbered			Odd-Numbered		Even-Numbered	
			Yugo (1)	Thai (2)	Isr (3)	Yugo (4)	Thai (5)	Isr (6)	Yugo (7)	Thai (8)	Yugo (9)	Thai (10)
1	6.87	.94		.56	.52	.59	.34	.33	.86	.32	.49	.21
2	6.25	.92			.35	.39	.36	.06	.45	.87	.43	.44
3	5.19	.81				.44	.43	.58	.01	-.15	.14	.14
4	6.95	1.05					.57	.55	.43	.18	.83	.35
5	6.23	1.06						.65	.15	.16	.24	.84
6	5.28	1.00							.04	-.25	-.01	.14
7										.47	.48	.16
8											.38	.33
9												.38
10												

The means, standard deviations, and Horst coefficients of the mean ratings on the Appraisal Guide are shown in Table 3. The Horst coefficients, averaging about .7, indicate that the mean ratings of the classes on the various items have substantial reliability; only three of the 27 rs are below .50.

Table 4 shows the zero-order and first-order partial correlations between the mean ratings of the Yugoslavia, Thailand, and Israel lessons on each Appraisal Guide item. Also shown are the rs between the self-reported attention scores for each of the three lessons. The zero-order rs between ratings of Yugoslavia and Thailand range from .60 to .83, except that for Item 7, which is .21. But the correlations between the mean ratings of these lessons with that of the Israel lesson are much lower, ranging from .00 to .36. The correlations show that students tend to rate their teacher similarly on the two lessons, but there is little or no tendency on their part to give the same rating to the taped lesson. For example, the correlation between the mean ratings of the Yugoslavia and Thailand lessons on Item 5, "Pacing the Lesson," is .62, which indicates that classes which rate the teacher high on the Yugoslavia lesson tend to rate the teacher high on the Thailand lesson for this item. The mean ratings of the Yugoslavia and Israel lessons on Item 5 correlated only .06, and the mean ratings on the Thailand and Israel lessons correlated only -.02. When the mean ratings on Yugoslavia and Thailand are adjusted for the mean ratings on Israel, the correlation remains the same, .62, indicating that a negligible amount of their variance was associated specifically with the ratings of the Israel lesson. This pattern of relationship prevails for all the Appraisal Guide items. Such results are reassuring in indicating that the correlations between the Yugoslavia and Thailand ratings do not arise merely from consistent class tendencies to rate more or less favorably.

The mean scores of the students on the Attention Report for the three lessons correlate substantially (rs = .68, .47, .48). The correlation between the mean scores on the Attention Report for the Yugoslavia and Thailand lessons, after the scores were adjusted for the mean score for attention to the Israel lesson, is .59. The latter r indicates that the level of attention attributable to the teachers is fairly consistent, even when variance due merely to consistent class tendencies, as measured by mean self-reported attention to the Israel lesson, has been partialled out.

Table 3

Means and Standard Deviations of Students' Mean Ratings
on each Appraisal Guide Item of the Yugoslavia, Thailand, and Israel Lessons

Appraisal Guide Items	Yugoslavia			Thailand			Israel		
	M	S. D.	Horst Coefficient	M	S. D.	Horst Coefficient	M	S. D.	Horst Coefficient
1. Clarity of Aims	3.69	0.55	.68	3.78	0.46	.55	2.96	0.60	.71
2. Organization of Lesson	4.07	0.50	.65	3.95	0.42	.69	3.37	0.57	.68
3. Beginning the Lesson	4.00	0.59	.36	3.83	0.59	.68	2.52	0.56	.56
4. Clarity of Presentation	4.26	0.53	.68	4.13	0.45	.71	3.04	0.56	.73
5. Pacing of Lesson	4.01	0.38	.82	3.88	0.40	.73	3.02	0.48	.63
6. Pupil Participation and Attention	4.29	0.74	.77	3.69	0.64	.75	2.05	0.54	.69
7. Ending the Lesson	3.54	0.77	.74	3.70	0.68	.52	2.87	0.56	.73
8. Teacher-Pupil Rapport	3.80	0.73	.67	3.67	0.66	.56	1.65	0.58	.70
9. Amount of Learning	3.56	0.59	.49	3.49	0.47	.43	2.38	0.58	.67
Mean Score on Attention Report	16.46	0.95		16.37	1.05		12.77	1.45	

Table 4

Generality of Ratings over Lessons: Correlations between Students' Mean Ratings of Three Lectures on Each Appraisal Guide Item and Mean Attention Scores, Unadjusted and Adjusted for Mean Response to Israel (N = 43 Classes Containing 10-31 Students)

Appraisal Guide Items	Unadjusted			Adjusted	
	Yugo vs. Thai	Yugo vs. Israel	Yugo vs. Israel	Yugo vs. Thai	Yugo Odd vs. Thai Even & Yugo Even vs. Thai Odd
					Mean of 2 rs between Mean Ratings, Correlated for Full "Length" a
	(1)	(2)	(3)	(4)	(5)
1. Clarity of Aims	.60	.34	.16	.59	.49
2. Organization of Lesson	.64	.28	.00	.66	.45
3. Beginning the Lesson	.70	.15	.21	.70	.52
4. Clarity of Presentation	.65	.17	.24	.64	.56
5. Pacing of Lesson	.62	.06	-.02	.62	.47
6. Pupil Participation and Attention	.69	.14	.20	.68	.55
7. Ending the Lesson	.21	.22	.06	.21	.14
8. Teacher-Pupil Rapport	.83	.15	.09	.83	.73
9. Amount of Learning	.71	.36	.15	.72	.56
Mean Score on Attention Report	.68	.47	.48	.59	.52

[a] This mean r was computed by (a) obtaining the two rs between mean ratings each based on half the number of students, (b) using the z-transformation to obtain the mean of the two rs, and (c) adjusting this mean r for increased "length," or the total number of students (Guilford, 1954, p. 407, Formula 14.39).

Corrected split-half estimates. The foregoing method, used to estimate the consistency of ratings and attention scores over lessons, employed an adjustment on the basis of ratings and attention scores on the Israel lesson. This method could be improved by eliminating additional irrelevant covariance due to students. The adjusted mean ratings by odd-numbered students on Yugoslavia were correlated with the adjusted mean ratings by even-numbered students on Thailand; similarly, the adjusted mean ratings by even-numbered students on Yugoslavia were correlated with the adjusted mean ratings by odd-numbered students on Thailand. The means of these two rs--corrected to obtain an estimate of the rs between means based on all students--are shown in column 5 of Table 4. These correlations are still substantial, but somewhat lower than the correlations in column 4 since more irrelevant covariance has been removed.

The correlations in Table 4 can be taken to indicate that student ratings of teachers and pupil attention over two different topics and two subsets of students are fairly consistent, even when the effects of consistent rating tendencies on the part of the students in various classes have been removed.

Correlations of Comprehension with Ratings and with Attention

The final question dealt with the correlations of the students' mean ratings and attention scores on the lessons with their mean comprehension scores on Yugoslavia, Thailand, and Israel. These correlations are shown in Table 5. Almost all the rs are substantial for all three lessons, indicating that teachers whose classes achieved higher scores on the comprehension test also tended to receive more favorable ratings on each Appraisal Guide item. These correlations remained substantial even when general tendencies to achieve highly and rate favorably, as measured by performance on and ratings of the Israel lecture, were adjusted for. Thus, the correlation between mean ratings on Item 9, "Amount of Learning," and the students' mean achievement test scores, or what they actually learned, on Yugoslavia is .61, on Thailand is .59, and on Israel is .66. The first two correlations drop only slightly, to .59 and .56, respectively, when both the mean ratings and the mean comprehension scores are adjusted for the mean ratings and mean comprehension scores on Israel. Hence the relationship beween what the students thought they learned and what they actually learned is not attributable to general attitude and achievement.

Table 5

Correlations of Comprehension with Ratings and Attention: Correlations between Mean Students' Ratings of Lectures and Self-Reported Attention Scores, Unadjusted and Adjusted, and Their Mean Comprehension Scores, Unadjusted and Adjusted
(N = 43 Classes Containing 10-31 Students)

Appraisal Guide Items	Correlation of Unadjusted Mean Rating with Unadjusted Mean Comprehension Score			Correlation of Adjusted Mean Rating with Adjusted Mean Comprehension Score		
	Yugoslavia	Thailand	Israel	Yugoslavia	Thailand	Mean of 4 rs Correlated for Full Length[a]
	(1)	(2)	(3)	(4)	(5)	(6)
1. Clarity of Aims	.64	.56	.47	.50	.39	.36
2. Organization of Lesson	.48	.48	.23	.24	.34	.27
3. Beginning the Lesson	.53	.41	.33	.46	.19	.28
4. Clarity of Presentation	.60	.44	.43	.53	.31	.35
5. Pacing of Lesson	.32	.28	.10	.21	.11	.17
6. Pupil Participation and Attention	.44	.23	.41	.47	.15	.31
7. Ending the Lesson	.24	.39	.34	.10	.35	.23
8. Teacher-Pupil Rapport	.46	.25	.02	.32	.10	.24
9. Amount of Learning	.61	.59	.66	.59	.56	.52
Self-Reported Attention Score	.59	.57	.52	.57	.55	.51

[a]Mean of the following four rs: (1) adjusted mean comprehension scores of odd-numbered pupils on Yugo vs. adjusted mean ratings by even-numbered pupils on Yugo; (2) odd-numbered mean comprehension scores on Thai vs. even-numbered mean ratings on Thai; (3) even-numbered mean comprehension scores on Yugo vs. odd-numbered mean ratings on Yugo; (4) even-numbered mean comprehension scores on Thai vs. odd-numbered mean ratings on Thai. Each r based on half of the students was adjusted for full "length" (see Guilford, 1954, p. 407, Formula 14.39) and then the four adjusted rs were averaged, using the z-transformation.

The mean ratings on Item 1, "Clarity of Aims," and Item 4, "Clarity of Presentation," correlate second and third most highly with the mean comprehension scores, in terms of the average r for both Yugoslavia and Thailand. The mean ratings on Item 5, "Pacing the Lession," and Item 8, "Teacher-Pupil Rapport," correlate least with the comprehension scores.

The correlations between self-reported attention scores and mean comprehension scores are much the same (r = .59, .57, and .52) for all three lectures. Adjustment by use of the scores for the Israel lesson has little effect on the other two correlations. This result again indicates that the correlation is not attributable to general tendencies to report attention and achieve comprehension in the same degree on all lessons.

Corrected cross-split estimates. The relationships of comprehension scores to mean ratings of the teacher and self-reported attention were also estimated by calculating the mean of four rs--namely, the rs between the following pairs of adjusted mean comprehension scores and adjusted mean ratings: (1) Yugoslavia comprehension of odd-numbered pupils vs. ratings by even-numbered pupils; (2) Yugoslavia comprehension of even-numbered pupils vs. ratings by odd-numbered pupils; (3) Thailand comprehension of odd-numbered pupils vs. ratings by even-numbered pupils; and (4) Thailand comprehension of even-numbered pupils vs. ratings by odd-numbered pupils (see column 6, Table 5). These correlations--corrected to obtain an estimate of the rs between means based on all students--show the relationship of achievement to ratings and attention when covariance due to involvement of the same students has been removed.

The most striking results in columns 4, 5, and 6 of Table 5 are the consistently positive correlations between the adjusted achievement scores and student ratings. The ratings have about the degree of correlation with achievement that should be expected for assessments of components of teaching performance. These correlations indicate that the specific behaviors defined by the Appraisal Guide are relevant to teacher effectiveness in explaining as measured by adjusted student achievement. The correlations are not so high, however, as to eliminate room for improvement by defining even more relevant behaviors. Some hints as to what these behaviors might be can be obtained from the results in Study II by Unruh.

Discussion

On the question of generality, the foregoing results warrant a positive conclusion. Generality was indicated by the correlation of .47 between the mean comprehension scores on one lesson and those on another, after adjustment for the mean ability of the students. When corrected split-half rs were obtained across topics (the adjusted means being based on random halves of the students), they averaged .41. Such generality indicates that the teacher's effectiveness in explaining does not depend entirely on the particular lesson being taught on a particular day to a particular group of students.

Beyond this finding, the present study demonstrated a method for estimating the degree to which various dimensions of the teacher's performance, as measured by mean students' ratings, correlate with the teacher's effectiveness, as measured by mean student achievement. The method entails adjusting both the mean rating and the mean achievement for student characteristics, as measured by their ratings of, and achievement on, a lesson taught without teacher variance, i.e., a tape-recorded lesson. The method also entails splitting each class into random halves to reduce irrelevant covariance due to basing two means on the same students. Estimated in this way, the rs between students' mean ratings and attention, on the one hand, and their comprehension, on the other, were positive and substantial, ranging from about .2 to .5.

II. THE MODALITY AND VALIDITY OF CUES TO LECTURE EFFECTIVENESS[3]

W. R. Unruh, University of Calgary

Teacher behavior can be recorded in various forms, or types of protocol: typewritten transcripts of lessons, audiotaped lessons, videotaped lessons, or combinations of these. Which of these kinds of record provides the basis for the most valid ratings of teacher effectiveness? Does the accumulation of cues, by involving additional sensory and perceptual modalities, improve the validity of judgments made about teachers?

Further, what are the correlates of effective explaining behavior? If it is found that raters can accurately judge teachers overall on a criterion of effectiveness, what are the bases on which their judgments are made? For example, are such rated characteristics as the teacher's warmth, vocal qualities, and pre-planning related to the teacher's effectiveness?

The investigation had three major parts: (a) an initial "postdiction rating study" to determine the relative effectiveness of seven kinds of protocols; (b) the "AV study"--similar to the initial postdiction rating study except that it was based on the use of audio-video protocols only; and (c) a study of correlates of effective explaining behavior as rated on a free-response instrument and a check-list of teacher characteristics.

Comparing the Protocols

In relation to the first question, seven types of protocol were compared in order to determine which type yielded the most accurate prediction of teacher effectiveness. The teachers' lectures were available in the form of typewritten transcripts, audiotapes, and videotapes, and the seven types of protocol were prepared from various combinations of these records.

The initial postdiction rating study. Four teacher-lessons were chosen randomly from each quarter of the distribution of adjusted mean comprehension scores on the Yugoslavia lesson, and four were similarly chosen from the Thailand group of lectures. Each of these lectures was then rated by eight twelfth-grade high school students randomly selected from a pool of 112 such students. Each group of eight raters was assigned to each of seven differ-

---

[3] A more complete report on this study is available in Unruh (1967).

ent protocols and to either of the two subject matters. Thus, different groups of eight judges each were assigned to the following seven kinds of protocols:

1.	Typewritten transcript only	T
2.	Audio record only	A
3.	Video record only	V
4.	Typewritten transcript plus audio record	TA
5.	Typewritten transcript plus video record	TV
6.	Audio record plus video record	AV
7.	Typewritten transcript plus video record plus audio record	TAV

Each student-rater read the article on which the lesson was based, and took the 10-item test based on that article. Then the raters were exposed to one of the seven types of protocols of the complete lessons of the four teachers in the subject-matter group to which they had been assigned. Following this they were exposed to a second six-minute portion of these lessons. Then they were asked to rate each teacher on a scale from one to 10 by postdicting the mean score they thought each teacher's class had obtained on the 10-item criterion test. The accuracy of this rating, when correlated with the actual mean scores made by the students of the four teachers, was used to evaluate the validity of the protocol.

The data derived from the initial postdiction rating study were analyzed in three different ways. First, rank-order correlation coefficients between the actual teacher ranks and the rater-assigned ranks were determined for each rater. The results of this analysis are presented in Table 6. The median rank-order correlation coefficient for the AV protocol group was .6 for the Yugoslavia sample and .7 for the Thailand sample. The median correlation coefficients for all other protocols were either negative or near zero.

Second, an analysis of variance was carried out to determine within each protocol (a) whether there was a significant difference among the ratings for each of the four teacher-lessons and (b) whether the mean ratings increased monotonically as the actual teacher scores increased. Variances significant at the .05 level were found among teacher-lessons for both content groups, and the mean ratings did increase monotonically as actual scores increased.

Table 6

Rank-Order Correlation Coefficients between Postdicted and Actual Ranks for Four Yugoslavia and Four Thailand Teacher-Lessons
N = 4 Protocols (Teacher-Lessons) per Rater

Protocol	RATERS 1	2	3	4	5	6	7	8	Median
(Yugoslavia)									
T	.0	-.6	-.6	-.6	-.6	-.8	-.8	-1.0	-.6
A	.8	.4	.4	.2	.0	.0	-.2	- .4	.1
V	.8	.8	.4	.2	.2	.2	-.2	- .4	.2
TA	.6	-.2	-.4	-.8	-.8	-.8	-.8	-1.0	-.8
TV	1.0	.8	.4	.2	-.4	-.4	-.4	- .8	-.1
AV	.8	.8	.6	.6	.6	.2	.0	- .8	.6
TAV	1.0	.8	.4	.4	.2	-.4	-.4	- .6	.3
(Thailand)									
T	.4	.2	.0	.0	-.4	-.4	-.4	- .4	-.2
A	1.0	.4	.4	.2	.0	.0	-.4	- .4	.1
V	1.0	.8	.4	.2	.2	.0	-.2	- .4	.2
TA	.4	.4	.2	.0	-.4	-.6	-.8	- .8	-.2
TV	.8	.4	.4	.2	.2	.2	-.4	- .8	.2
AV	.8	.8	.8	.8	.6	.4	.4	- .6	.7
TAV	.8	.4	.4	.2	.0	-.2	-.4	- .4	.1

Third, accuracy-of-rating scores were computed as the sum of squared differences between the actual teacher ranks and the rater-assigned ranks, and these scores were subjected to analysis of variance. Differences in accuracy among protocols significant at the .01 level were found for Yugoslavia, and at the .08 level for Thailand. Tests of paired means, however, showed that AV was significantly different ($p<.01$) only from T and TA in the Yugoslavia group and only from T ($p<.10$) in the Thailand group. In view of the findings noted above, and because the accuracy of ratings based on the AV protocol was highest in both subject-matter groups, it appeared reasonable to conclude that the AV protocol was the best basis on which to collect further data.

The AV study. An additional 60 judges (30 assigned randomly to the four lessons on Yugoslavia, and 30 to the four Thailand lessons) then rated the AV protocols. The ratings by these 30 judges were pooled with those of the eight judges used in the two AV groups in the initial rating study, thus providing responses from a total of 38 raters in each AV group.

These ratings were analyzed in the same way as those in the initial postdiction rating study, except that no comparison between protocols was made. The median rank-order correlations between actual and postdicted ranks were .56 and .55 for the Yugoslavia and Thailand samples, respectively. F-ratios significant beyond the .001 level were obtained for differences in mean rating assigned each of the four teachers, and the postdicted means increased monotonically as the actual means increased, with the exception that Teachers 1 and 2 of the Yugoslavia sample and Teachers 2 and 3 of the Thailand sample were reversed. Apparently, the twelfth-grade students who served as judges could rate the teachers with reasonable validity.

Correlates of Explaining Effectiveness

The second purpose of the study was to determine correlates of effective explaining as perceived by the judges. In view of the findings noted above, only the data derived from the AV protocols were analyzed in this phase of the study.

The free-response study. After the judges had supplied their overall rating for each lesson, they were asked to write on a blank form at least six adjectives or phrases describing strengths or weaknesses of the teacher,

the lesson, or the way in which the lesson was presented. This procedure was used to determine the bases on which the raters had made their judgments.

A content analysis of the resulting 1,768 responses made it possible to assign responses either to one of 17 positive categories or to one of 18 negative categories.

The discriminatory value of each category was determined according to whether it showed a significant difference among teacher-lessons in the frequency of responses, and whether there was a substantial correlation between these frequencies and the actual teacher scores. Whether the first requirement was met was tested by means of the chi-square one-sample test, as indicated in Tables 7 and 8. Whether the second requirement was met was determined in terms of rs between the frequency of rater responses and the actual teacher scores, also presented in these tables. In view of the exploratory nature of this study, it was decided that a category would be accepted as having discriminatory value if the chi-square test showed differences in frequency of a free-response category significant at the .05 level, and if the correlation between free-response frequencies and actual teacher scores was above $\pm$ .40.

Using these bases for choosing categories for discussion, and looking at the Yugoslavia and Thailand responses separately, one can discern the following general picture. The positive free-responses of student-raters in the Yugoslavia group portray the good teacher as one who:

1. was organized and had planned well,
2. spoke at an appropriate cognitive level,
3. was serious and did not openly display a sense of humor,
4. had and used an outline effectively, and
5. had a good introduction in the sense that he stated objectives clearly and provided adequate background information.

In the Thailand sample the good teacher was seen as one who:

1. covered the material well,
2. made effective use of visual aids,
3. knew and understood the subject matter, and
4. was interesting and able to keep up the interest of the class.

Table 7

Chi-Square Tests of Significance and Pearson Product-Moment Correlation Coefficients for Positive Free-Response Items

Category Name	Yugoslavia χ^2	Yugoslavia r^a	Thailand χ^2	Thailand r^a
1. Voice and Speech Qualities	.04	-.57	3.76	.99
2. Relevance and Emphasis	10.46*	-.19	7.60	.79
3. Coverage	3.43	.98	9.86*	.52
4. Achievement of Aims	5.73	.04	.81	.78
5. Gesture and Movement	.33	.54	5.20	.64
6. Use of Visual Aids	7.71	-.53	9.36*	.51
7. Organization and Planning	24.85**	.95	2.88	.88
8. Knowledge of Material	7.23	.72	12.87**	.86
9. Level of Speech or Vocabulary	10.57*	.69	3.92	.93
10A. Sense of Humor	8.40*	-.45	25.73**	-.17
B. Enthusiasm, Vitality, Energy, etc.	3.47	-.70	22.80**	.25
C. Interest and Involvement in the Topic	3.00	-.20	24.00**	.27
D. Confidence, Poise, etc.	6.07	-.44	11.33**	-.20
E. Friendliness, Warmth, Casual Manner, etc.	3.88	-.48	17.00**	-.14
11. Interesting	2.25	-.56	10.42*	.78
12. Rate of Delivery	1.56	-.31	.00	.00
13. Use of Outline	28.66**	.97	7.16	.26
14. Use of Examples and Illustrations	.50	-.38	2.33	-.79
15. Good Introduction	10.57*	.41	.25	-.03
16. Good Conclusion	3.33	.60	3.80	.70
17. Unclassified	2.86	.37	2.21	.94

* $p<.05$

** $p<.01$

[a] N = 4 Teacher-Lessons

Table 8

Chi-Square Tests of Significance and Pearson Product-Moment Correlation Coefficients for Negative Free-Response Items

Category Name	Yugoslavia χ^2	Yugoslavia r^a	Thailand χ^2	Thailand r^a
1. Voice and Speech Qualities	2.89	-.19	3.45	-.41
2. Relevance and Emphasis	1.31	-.74	3.73	-.67
3. Coverage	.00	.00	6.44	-.45
4. Achievement of Aims	5.00	.09	10.00*	-.64
5. Gesture and Movement	.78	.81	7.45	.68
6. Use of Visual Aids	5.87	-.93	6.00	.31
7. Organization and Planning	14.15**	-.82	12.68**	-.89
8. Knowledge of Material	25.24**	-.93	19.66**	-.88
9. Level of Speech or Vocabulary	15.02**	-.40	2.57	.89
10A. Lacks Sense of Humor	2.00	.75	+	.28
B. Lacks Enthusiasm, Vitality, Energy, etc.	5.66	-.06	9.73*	-.75
C. Lacks Interest and Involvement in the Topic	7.54	.82	2.20	-.68
D. Lacks Confidence, Poise, etc.	10.48*	-.98	11.94**	.66
E. Lack of Friendliness, Warmth, Casual Manner, etc.	3.88	.08	5.00	-.71
11. Uninteresting, Boring	11.00*	.42	33.23**	-.83
12. Rate of Delivery	.38	.31	12.40**	.08
13. Use of Outline	+	-.73	+	-.03
14. Use of Examples and Illustrations	+	.00	+	.00
15. Poor Introduction	+	.54	+	.00
16. Poor Conclusion	+	.80	6.80	-.86
17. Too Many "Uh's"	3.60	-.68	+	-.79
18. Unclassified	3.14	-.52	1.06	-.35

* $p<.05$

** $p<.01$

[a] N = 4 Teacher-Lessons

Note: Where the number of responses to an item was less than two, no χ^2 difference could be determined. This is indicated in the table by a (+).

Similarly, the negative free-responses of the Yugoslavia group portray the poor teacher as one who:

1. did not plan well and was not well organized,
2. did not know the subject matter well,
3. did not speak at an appropriate cognitive level,
4. lacked confidence and poise, and
5. was not interesting or was unable to keep up class interest.

In the Thailand group a similar picture emerged. (Negative Categories 4, 7, 8, 10B, 10D, and 11 met the requirements for discriminative value.) It should be noted, however, that the sign of the correlation coefficient for Category 10D is reversed so that, whereas the Yugoslavia group saw the poor teacher as lacking confidence and poise, this was not true of the Thailand group. Similarly, there was a reversal in Category 11, where the raters indicated that the poor teacher was more often seen as boring than the good teacher. Thus the poor teacher was described by these raters as one who:

1. was unable to present the material in a clear way,
2. had not planned well and was not well organized,
3. did not know the subject matter well,
4. lacked enthusiasm and vitality,
5. was confident and poised to a greater extent than the better teacher, and
6. was boring and unable to keep up the interest of the class.

Positive Categories 3, 5, 7, 8, 9, and 16, although not all of them discriminated among the four teacher-lessons on the basis of the chi-square tests, and therefore not all of them fully met the requirements for validity stated above, nevertheless correlated highly with the criterion scores across both teacher-lesson samples and may have some value in discriminating between good and poor lecturers.

Taken as a whole, the <u>consistent</u> results for the free-response categories suggest several conclusions which may be useful in further teacher effectiveness studies. First, as far as these judges are concerned, the most important aspects of good teaching appear to involve teacher activities

related primarily to preparation and presentation. Thus they rate highest those teachers who plan well, are well organized and prepared, speak at an appropriate cognitive level, use an outline--which may be related to planning--and cover the material well. These teacher activities are basically cognitive in nature. That is, they involve mainly matters which describe the teachers' pre-planning and application of strategies aimed at structuring lesson materials so as to make the subject matter meaningful. Even the references to poise and self-confidence, though not consistently related to good teaching in this study, can be construed as a sign that the teacher is sure of his material and presentation. It may be, of course, that many of the responses made by the raters are based on the appearance of the teacher--that is, he may not really be better prepared or organized, but he appears to be sure and presents his material in a business-like and efficient manner.

Second, items relating to personality variables and vocal quality do not discriminate consistently in this context. According to the positive free-responses made by the Yugoslavia raters to Category 10A, the better teacher is serious rather than humorous. The only other reference to what may be non-cognitive aspects of the presentation are the references to enthusiasm and vitality (Category 10D of the Thailand group), and the references to Negative Category 11, which describe the good teachers as boring in the Yugoslavia sample and the poor teachers as boring in the Thailand sample. This reversal is difficult to explain. It may be that this factor does not discriminate between good and poor teachers, or it may result from the raters' viewing a business-like presentation as boring but still giving a high general rating to the teacher because of other positive characteristics.

The check-list study. In this phase of the study, the judges were exposed once more to each of the lessons for a six-minute period and were asked to rate each teacher on a series of 27 seven-point bi-polar scales consisting of adjectives and phrases selected from the research literature because of their presumed relevance to teacher effectiveness.

An analysis of variance was carried out for each of the 27 scales. Pearson rs between scale ratings and actual teacher scores were also computed.

The results of these analyses appear in Tables 9 and 10 and indicate that the descriptions of good and poor explainers, by means of these scales, agreed in general with those provided by the raters' free-responses. Scales judged

Table 9

F-Ratios and Levels of Significance for 27 Seven-Item Scales for Four Yugoslavia and Four Thailand Teacher-Lessons

Scale No.	Scale Description	Yugoslavia F-ratio	Thailand F-ratio
1.	Businesslike vs. slipshod	12.87**	25.99**
2.	Clear vs. obscure, vague	7.97**	14.13**
3.	Dynamic vs. phlegmatic	12.63**	2.87*
4.	Emphatic vs. unemphatic	12.28**	1.93
5.	Enthusiastic vs. unenthusiastic	17.97**	6.25**
6.	Energetic vs. lethargic	30.08**	10.57**
7.	Friendly vs. unfriendly, aloof	27.13**	6.50**
8.	Fluent in expression vs. halting in expression	3.49*	8.39**
9.	Humorous vs. dull	31.97**	4.94**
10.	Interesting vs. boring	13.02**	6.12**
11.	Imaginative vs. unimaginative	18.43**	2.53
12.	Interested vs. uninterested	12.21**	2.35
13.	Poised vs. awkward	1.29	11.33**
14.	Positive attitude vs. negative attitude	6.35**	1.59
15.	Stimulating vs. dull, unstimulating	10.38**	2.56
16.	Skillful vs. inept, unskillful	15.78**	11.88**
17.	Warm vs. cold	13.75**	4.68**
18.	Knows and understands subject vs. does not know and understand subject	10.56**	13.42**
19.	Lesson is well planned vs. lesson is not well planned	11.11**	14.45**

Table 9 (continued)

Scale No.	Scale Description	Yugoslavia F-ratio	Thailand F-ratio
20.	English expression good vs. English expression not good	.61	14.74**
21.	States objectives of lesson clearly vs. does not state objectives of lesson clearly	2.06	7.45**
22.	Makes relationships clear vs. does not make relationships clear	3.70*	4.09*
23.	Clearly indicates when moving from one topic to another vs. does not clearly indicate when moving from one topic to another	2.75*	3.64*
24.	Makes effective use of voice vs. does not make effective use of voice	9.03**	2.34
25.	Points out clearly what should be learned vs. does not point out clearly what should be learned	4.03**	2.98*
26.	Gives adequate amount of detail vs. does not give adequate amount of detail	1.46	2.07
27.	Summarizes and reviews frequently vs. does not summarize and review frequently	1.47	2.37

* $p<.05$

** $p<.01$

Table 10

Pearson Product-Moment Correlation Coefficients and Estimates of Monotonic Relationships for Four Yugoslavia and Four Thailand Teacher-Lessons

Scale No.	Scale Description	Yugoslavia N = 38 r	Mon.[a]	Thailand N = 38 r	Mon.[a]
1.	Businesslike vs. slipshod	.20	+	.55**	+
2.	Clear vs. obscure, vague	.35*	+	.45**	+
3.	Dynamic vs. phlegmatic	.32*	+	-.18	-
4.	Emphatic vs. unemphatic	.32*	+	-.15	-
5.	Enthusiastic vs. unenthusiastic	.35*	+	-.33*	-
6.	Energetic vs. lethargic	.42**	+	-.35*	-
7.	Friendly vs. unfriendly, aloof	.07	-	-.28	-
8.	Fluent in expression vs. halting in expression	.14	+	.39*	+
9.	Humorous vs. dull	.05	-	-.20	-
10.	Interesting vs. boring	.29	+	.02	-
11.	Imaginative vs. unimaginative	.22	-	.03	-
12.	Interested vs. uninterested	.27	+	-.16	-
13.	Poised vs. awkward	-.03	-	.43**	+
14.	Positive attitude vs. negative attitude	.21	+	.05	-
15.	Stimulating vs. dull, unstimulating	.24	+	-.02	-
16.	Skillful vs. inept, unskillful	.40**	+	.43**	+
17.	Warm vs. cold	.04	-	-.19	-

Table 10 (continued)

Scale No.	Scale Description	Yugoslavia N = 38 r	Mon.[a]	Thailand N = 38 r	Mon.[a]
18.	Knows and understands subject vs. does not know and understand subject	.28	+	.44**	+
19.	Lesson is well planned vs. lesson is not well planned	.26	+	.49**	+
20.	English expression good vs. English expression not good	.05	+	.47**	+
21.	States objectives of lesson clearly vs. does not state objectives of lesson clearly	.06	+	.35*	+
22.	Makes relationships clear vs. does not make relationship clear	.21	+	.25	+
23.	Clearly indicates when moving from one topic to another vs. does not clearly indicate when moving from one topic to another	-.02	-	.23	+
24.	Makes effective use of voice vs. does not make effective use of voice	.32*	+	-.02	-
25.	Points out clearly what should be learned vs. does not point out clearly what should be learned	.02	+	.21	+
26.	Gives adequate amount of detail vs. does not give adequate amount of detail	.09	+	.16	+
27.	Summarizes and reviews frequently vs. does not summarize and review frequently	.05	-	.01	-
		Mult R = .69		Mult R = .74	

* $p<.05$

** $p<.01$

[a]Monotonic relationship as described above is indicated here by a (+). A (-) indicates that such a relationship was not found with reference to the means of the rater-assigned scores.

consistently valid by the procedures described above for both teacher-lesson samples indicated that the good teachers were skillful in presenting the material, made the content of the lesson clear, knew the subject well, and had apparently done a good job of planning the lesson. The opposite was true of the poor teachers.

Again, these findings suggest that the good teacher is seen by these raters as the one who is verbally and cognitively in control of the situation. No relationships consistent across content groups were found between non-cognitive variables and the achievement criterion.

Implications

The findings of this study suggest ideas and questions on which to base further research.

1. The results of this study indicated that the AV protocol was the best source of cues for rating the teacher's effectiveness in explaining when such effectiveness was measured on the basis of student achievement. One reason for this finding may have been that the AV presentation was the one most similar to the type of classroom presentation with which the judges were familiar and therefore capable of evaluating. This finding suggests that in studies where ratings of teacher effectiveness are to be made, AV records should be used in preference to other records of behavior, such as audiotape recordings and typewritten transcripts. Yet it should be noted that the ratings made were overall effectiveness ratings, and the researcher, in deciding which record of behavior to use, would need to decide whether the variables with which he is concerned could best be measured via such a record. In other words, which protocol is best may depend on the variable being investigated.

2. The curriculum materials used in this study were more or less typical of the social studies; research is needed to determine whether the characteristics which discriminated among good and poor teachers in this study would also hold for other subject matters. Similarly, research at different grade levels is needed. The effects of various teacher characteristics may interact with subject matter and grade level.

3. The dependent variable against which significant characteristics of explaining behavior were reflected in this study was the adjusted mean achieve-

ment scores of the classes of the teachers in each sample. While this was probably a satisfactory first approximation for establishing a criterion, further studies might do well to test this criterion more extensively. One approach might be to play the videotaped lessons of a group of teachers to a number of groups of pupils assigned at random to each teacher-lesson. Pupils, especially at the higher grade levels, may be able to adjust to certain aspects of their own teacher's behavior. But playbacks to pupils who had never before been exposed to the teacher would help to control for such accommodation and thus enhance the validity of the criterion measure.

4. Further research might also be directed at refining the dependent variable used in this study. "Explaining," used here as a "micro-criterion" of effectiveness, is still a broad criterion. It may be desirable to break down the explaining process into smaller units and to measure the effectiveness of teachers in handling these smaller aspects of the task.

5. The perceived characteristics of teachers shown in this study to be related to explaining effectiveness are similarly broad in nature. These characteristics should be defined operationally in behavioral terms and then re-validated.

III. OBJECTIVELY MEASURED BEHAVIORAL PREDICTORS OF EFFECTIVENESS IN EXPLAINING[4]

Barak Rosenshine, Temple University

This investigation was aimed at determining objectively measured teacher behaviors that discriminate between more and less successful explanations of social studies material.

The variables investigated were the stimuli received by the pupils, that is, the verbal and non-verbal behaviors of teachers while they lectured. For some aspects of the lectures, a grammar of objective terms already existed which could be used to categorize the characteristics of the lectures. For example, a number of existing grammars can be used to categorize the length and type of individual words or independent clauses. But, for most of the aspects of the lectures, it was necessary to construct an analytic grammar by selecting significant variables which had been developed in other kinds of investigations and adapting them to the process of teaching and giving explanations by lecture.

The variables were developed from 27 categories derived from the research in four areas: linguistics, instructional set, experimental studies of instruction, and multivariate studies of the behavioral correlates of teaching effectiveness.

Thirty lectures were selected for investigation. Five high-scoring and five low-scoring lectures, as measured by the students' residual mean achievement scores on the test on Yugoslavia, comprised a "hypothesis group." Another five high-scoring and five low-scoring lectures on Yugoslavia comprised a "validation group." The five most and the five least effective teachers in explaining the material on Thailand were used as a "cross-validation group." The frequency of occurrence of each variable was first tabulated using the lectures in the hypothesis group. Those variables that discriminated at the .15 level between the high-scoring and low-scoring lectures in the hypothesis group were then counted in the remaining two groups. The significance of the discrimination was tested by means of a two-way analysis of variance in which the hypothesis, validation, and cross-validation groups formed the columns,

---

[4] A more complete report on this study is available in Rosenshine (1968).

and the high-scoring and low-scoring lectures formed the rows.

Results

Linguistic categories. The linguistic categories were developed primarily from the research on readability. The most frequently consulted references were the summaries of research by Chall (1958) and Klare (1963) and the factor analytic studies by Brinton and Danielson (1958) and Stolurow and Newman (1959).

The 43 variables investigated in this area were developed from nine categories:

1. Word length
2. Total number of relevant words
3. Length and structure of independent clause units
4. Prepositional phrases
5. Readability Estimate--the readability of the lectures as determined by the multiple-regression formula developed by Flesch (1948)
6. Personal references--counts of first and second person pronouns
7. Negative sentences--counts of sentences containing not modifying the verb, not modifying a noun, and/or not only or an equivalent phrase
8. Passive verbs--counts of independent and/or dependent clauses containing passive verbs
9. Awkward and fragmented sentences--counts of sentences which depart from usual sentence construction and/or phrases which lack a subject or a verb but which add information (i.e., "Now to foreign affairs").

Variables in four of these nine categories discriminated between the high-scoring and low-scoring lectures in the hypothesis group, but none of the differences in frequencies was significant across the three groups. For the hypothesis group only, the high-scoring lectures contained fewer syllables per word, independent clause units with more words and clauses, and more prepositional phrases. They also contained more words rated as directly or indirectly relevant to the criterion questions.

Instructional set. The next two categories were developed from the experimental research on the influence of pre-instructional procedures, or instructional set, upon the effects of subsequent presentations. The most frequently consulted references were the research of Ausubel (1963), Hovland, Lumsdaine, and Sheffield (1949), May and Lumsdaine (1963), Allen (1955), and Rothkopf (1966).

Thirty-seven variables in two categories were investigated. One category, (10) structuring sets, contained variables which might resemble "discriminating advance organizers," that is, words or phrases which indicate that the speaker is attempting to clarify distinctions between new and previously learned material. The other category, (11) focussing or arousing sets, contained variables which might identify phrases designed to arouse or focus attention. None of the variables in these two categories discriminated between the high and low lectures in the hypothesis group.

Presentational categories. Nine "presentational" categories were developed from a broad class of experimentally tested variables which might be related to explaining ability. The most frequently consulted references were reviews by Travers, et al. (1964), Lumsdaine (1963), and Petrie (1963). The nine categories were: (12) use of rule and example pattern, (13) number of examples, (14) organization of topics, (15) use of enumeration, (16) movement and gesture, (17) breaks in speech, (18) use of maps and chalkboard, (19) rate of speech, and (20) repetition and redundancy.

Variables in three of these categories discriminated between the high and low lectures in the hypothesis group, but not across the three groups. In the hypothesis group only, the high teachers spoke faster, used fewer pauses and verbalized breaks, used the chalkboard less frequently, and used maps more often. The high teachers also appeared to use more between-sentence repetition, but this category was dropped because it was impossible to develop a reliable coding system for repetition or review.

Variables in two categories--rule and example pattern, and movement and gesture-discriminated between the high and low lectures across the three groups. These findings will be discussed below.

Multivariate studies of teaching behaviors. The categories studied in this fourth area were developed from research on the relationship between spe-

cific teaching behaviors and measures of adjusted student gain. Representative studies are those by Medley and Mitzel (1959), Flanders (1965), Spaulding (1963), Bellack, et al. (1966), and Soar (1966).

Five categories based upon the significant findings of these investigators were selected for investigation, but variables in none of these categories appeared in sufficient frequency to be counted. The categories were: (21) verbal hostility, (22) non-verbal affect, (23) reference to pupils' interests, (24) expansion of pupils' ideas, and (25) ratio of acceptance and praise to criticism.

The frequency of occurrence of variables in two additional categories was counted: (26) conditional words and (27) explaining links. The frequency of conditional words such as "but," "however," and "although" did not discriminate between the high and low lectures in the hypothesis group. The frequency of explaining links was significant across the three groups and will be discussed below.

Discussion

Variables in 21 of the 27 categories occurred with sufficient frequency to merit counting and could be reliably counted. Variables in 10 of these categories discriminated between the high and low lectures in the hypothesis group, and variables in three of these 10 categories discriminated between the high and low lectures across all three groups. The latter three, consistently discriminating, categories, as shown in Table 11, were rule and example pattern, explaining links, and gesture and movement.

Gesture and movement. Gesture was defined as movement of the arms, head, or trunk, and movement was defined as lateral (left and right) movement of the teacher from one fixed place to another. When the unit of measure was per lecture, the high groups had more ($p<.05$) gestures, seconds of gesture, movements, and seconds of movement than the low groups. Three of these variables were significant or nearly significant when the unit was per minute, and one of these variables (movement) was nearly significant ($p<.06$) when the unit of measure was per one hundred words.

These gestures and movements may have the effect of arousing or focussing attention. However, verbal variables taken singly and in combinations which might have been classified as attention-arousing variables did not discriminate

Table 11

The Discriminating Power of Variables within 21 Categories

Area	Category
Linguistic Categories	*1. Word length
	*2. Word relevance
	*3. Independent clause length and structure
	*4. Prepositional phrases
	5. Readability estimate
	6. Personal references
	7. Negative sentences
	8. Passive verbs
	9. Awkward and fragmented sentences
Instructional Set	10. Structuring sets
	11. Focussing or arousing sets
Presentational Categories	**12. Rule and example pattern
	13. Number of examples
	14. Organization of topics
	15. Use of enumeration
	**16. Gesture and movement
	*17. Breaks in speech
	*18. Use of map and chalkboard
	*19. Rate of speech
	20. Repetition and redundancy
Multivariate Studies	21. Verbal hostility
	22. Non-verbal affect
	23. Reference to pupils' interests
	24. Expansion of pupils' ideas
	25. Ratio of acceptance and praise to criticism
	**26. Explaining links
	27. Conditional words

*Variables in this category discriminated between high and low lectures in the hypothesis group, but not across the three groups.

**Variables in this category discriminated between the high and low lectures across the three groups.

between the high and low lectures in the hypothesis group. These verbal variables included phrases stating the importance of material or recalling material, cognitive reversal, and references to problems and conflict.

Rule-example-rule pattern of discourse. The term rule refers to the use of a summary statement before or after a series of examples. The high lectures contained a higher frequency ($p<.01$) and percentage ($p<.01$) of patterns which contained two rules, one before and one after a series of examples. The low lectures had a higher frequency ($p<.01$) and percentage ($p<.01$) of patterns with a single rule only before the examples; the low lectures also had a significantly higher frequency ($p<.01$) and percentage ($p<.01$) of sequences with a single rule given either before or after the examples.

These results may indicate that a pattern of explanation which opens with a structuring statement, follows it with details, and concludes by re-stating the structuring statement is more effective than other patterns. But it is difficult to generalize this finding to the analysis of written or spoken prose because it is difficult to distinguish between examples and statements of fact. In this study, we were able to identify examples only by referring to the original article and selecting as examples those state-ments of fact which were preceded or followed by an organizing principle. If we had not the original articles as references, the coding results might have been much less consistent.

An extension of this idea might be the proposition that a paragraph would be more effective if it began and ended with a topic sentence. But we were unable to identify topic sentences in the lectures studied in this investiga-tion.

Explaining links. The concept of a cognitive process labeled explaining was developed from the research of Bellack, et al. (1966), who developed their work in this area from the research of Smith and Meux (n.d.). The explaining process was defined by Bellack as consisting of statements describing the re-lation between objects, events, or principles, or statements reporting either cause and effect or comparison and contrast. Words such as "because" and questions containing "why" were cited by Bellack, et al., as indicators of explanation.

In this investigation the frequency of explanations was assessed by counting explaining links, that is, prepositions and conjunctions which indicated the cause, result, means, or purposes of an event or idea. Words and phrases such as "because," "in order to," "if...then," "therefore," and "consequently," were counted, as well as specified instances of words and phrases such as "since," "by," and "through." The high lectures contained more ($p<.01$) of these words in each of three units of measure: per lecture, per minute, and per hundred words.

Words such as these explaining links may function to link phrases either within or between sentences so that a phrase or clause containing an explaining link elaborates and expands upon another phrase or clause. This special linkage may be illustrated by the following three sentences which are almost identical:

1. The Chinese dominate Bangkok's economy, and they are a threat.

2. The Chinese dominate Bangkok's economy, but they are a threat.

3. The Chinese dominate Bangkok's economy; therefore they are a threat.

The third sentence may be the easiest to comprehend because it contains the explaining link "therefore" instead of other words such as "and" or "but." Different types of explaining links also seem to be interchangeable, as in the following three examples:

1. The Chinese dominate Bangkok's economy; therefore, they are a threat. (Statement of consequence)

2. The Chinese are a threat because they dominate Bangkok's economy. (Statement of cause)

3. By dominating Bangkok's economy, the Chinese are a threat. (Statement of means)

It should be noted that the explaining links counted in this study were only a convenience for identifying "explaining sentences." There is no claim that the words selected represent all words which could be selected. This category should be investigated more closely, eliminating words which are not true explaining links and determining whether certain nouns and verbs can be included in this category.

Semantic subordination. The notion that an explaining link introduces a clause which adds to or elaborates upon another clause is close to the grammarian's definition of subordination. But in this investigation, common measures of subordination such as dependent clauses and prepositional phrases did not discriminate between the high and low lectures. The words chosen as explaining links included subordinating and coordinating conjunctions, as well as certain adverbs and prepositions. Although these words are grammatically dissimilar, they are semantically similar; they introduce a clause or phrase which states a means, reason, or consequence for the main clause. It might be productive in future studies to analyze teachers' statements by this semantic method rather than by traditional sentence structure.

There are many other phrases or words which are grammatically dissimilar but perform the same semantic function. The prepositional phrase in general does not appear to be different from the adverb generally; the prepositional phrase at the present time appears similar to the noun today; the dependent clause if a person were to visit Thailand is similar to the phrase a person visiting Thailand.

The converse may also be true. Some forms of speech are grammatically similar but perform different semantic functions. For example, prepositional phrases can introduce a major topic or minor topic, a definition, or a summary; and they can be used for sequencing, emphasizing, or elaborating. Dependent clauses and participial phrases can also perform these functions. Because of this variety in function, an indiscriminate increase or decrease in the proportion of certain grammatical structures, such as prepositional phrases or gerunds, in a communication cannot be expected to affect comprehension.

The results from the coding of explaining links suggest that only some of the functions of subordinate clauses and phrases are effective in increasing the comprehensibility of communications. For this reason, attempts to discriminate between effective and ineffective lectures by counting different parts of speech may have limited promise. More significant results may be obtained when subordination is investigated by distinguishing certain subordinate clauses and phrases from others according to the way in which they function in a sentence. The present investigation of explaining links may have identified one type of functional subordination.

Recommendations for Experimental Studies

There is as yet no firm basis for translating any of these findings into recommendations for teaching because the investigation was not an experimental study. A post hoc study such as this one can only suggest potential correlates of teaching effectiveness. Experimental research will be necessary before we can claim that any of the presently significant variables represent causal factors.

One such experiment could proceed by selecting teachers whose performance was low as measured both by the adjusted achievement scores of their students and by the frequency of their use of movement and gesture, the rule-example-rule pattern, and explaining links. Some of these teachers would then be trained to use these behaviors more frequently. These teachers would then teach new material to new classes, and their effectiveness in explaining would be compared with that of a similar group of untrained teachers.

A second experiment could test whether the verbal and non-verbal findings are independent of each other. Some students would read the transcripts of the high-scoring lectures and other students would read the low-scoring lectures. The ranking of these students' achievement scores could be compared with those of the students in the original study. Using transcripts alone would eliminate the effects of non-verbal variables such as movement, gesture, and rate. Such procedures might also control for factors such as quality of voice and teacher personality.

Another way of studying the effects of the use of explaining links and the rule-example-rule pattern would be to add or eliminate instances of these behaviors from the lectures. Several transcripts from this investigation could be used, some high-scoring and some low-scoring. The high-scoring lectures could be altered so that the explaining links and the second statement of the rule are removed, and the low-scoring lectures could be altered by adding explaining links and a second statement of a rule to the original material. There would then be four lectures, an original and an altered version for both the high and the low lectures. Experiments could test hypotheses that the manipulations decrease the effectiveness of the high-scoring lecture and increase the effectiveness of the low-scoring lecture.

If the significant findings concerning explaining links and the rule-example-rule pattern are replicated through the use of transcripts alone,

then these variables may be considered useful additions to the correlates of readability. Consideration of these variables may explain some of the inconsistent findings in studies of instructional set. The use of instructional sets may decrease in effectiveness as the number of explaining links in the instructional material is increased. If so, explaining links may provide the same sort of linkage and organization within the lecture as the instructional set gives in the introduction to the lecture.

IV. COMPUTER ANALYSIS OF TEACHERS' EXPLANATIONS[5]

Daryl Dell, Stanford Research Institute
Jack E. Hiller, United States Army

This study was an exploratory effort to apply computer programs developed in other areas to educational problems and to develop new variables whose frequency of occurrence might correlate with effectiveness in explaining. It was possible to test the latter variables using specially developed computer programs.

Uses of the Computer

Reliability is always a problem in content analysis studies of the type reported above. The problem arises particularly when human judges are used for such boring or complex tasks as counting word length, sentence length, the frequency of occurrence of different words, and words used in various functions. Page (1966) has demonstrated that the computer can be used with high scoring reliability and high objectivity to tally items such as average sentence length, average word length, number of commas, standard deviation of sentence and word length, word frequency, and word ratios. The report on the authorship of the Federalist Papers (Mosteller and Wallace, 1964) presents many notable examples of this technique.

Other investigators have developed computer programs which go beyond the counting of word length or word frequency. For example, the General Inquirer (Stone, et al., 1966) includes a dictionary look-up system which counts the frequency of different types of words. Systems such as these include the use of a "dictionary" of words classified as to type, affect loading, or any other dimension chosen by the investigator. The computer then is given textual material, and it uses the dictionary to report on the frequencies of designated words in this material. The dictionaries can be changed to meet the needs and growing knowledge of the investigators, and the original data can be rerun against the new hypotheses which were used to develop the new dictionaries.

Initial Procedures in This Study

Transcripts of selected lectures of the teachers in this study were key-

---

[5]For a more complete report, see Hiller (1968).

punched on IBM cards, and were used as the material to be analyzed. The first approach was to use the computer programs developed by Page and his staff to count items such as average sentence length, average word length, number of commas, and the standard deviation of sentence and word length. The scores of the lectures on these variables did not correlate significantly with the mean adjusted achievement scores for the lectures. One of the difficulties may have been the questionable reliability of inserting commas and periods in transcribed spoken prose. At any rate, because of these difficulties and because the results appeared to confirm those reported by Rosenshine in Study II of this report, this approach was discontinued.

Selection and Development of Dictionaries

The General Inquirer could not be used to count the frequencies of certain types of words because the program was not useable with the computer equipment available. But the use of a dictionary to count the frequency of words classified according to semantic and affective categories seemed promising, and this approach was used in the study. Three existing dictionaries were used, and three additional dictionaries were created to fit our purposes.

The first existing dictionary was the Stanford Political Dictionary developed by Holsti (1966). This dictionary contains over 3,000 words that have been evaluated on the semantic differential scales developed by Osgood, Suci, and Tannenbaum (1957). These scales consist of three bi-polar dimensions yielding the following six categories: good or bad, active or passive, and strong or weak. Within each of the six categories the words in the dictionary are rated on a three-point continuum. Each of the 3,000 words in the dictionary was rated only on the scales that were judged to be appropriate to it. The program as developed by Holsti and his group required the use of specially coded data, and evaluated the data in terms of perceptual viewpoint.

Two additional existing dictionaries were used. These were the 83 categories of the Harvard Third Psychosociological Dictionary (Stone, et al., 1966) and the Dale list of the 3,000 most commonly used words (1948).

Three additional dictionaries were developed specifically for this project. Hiller developed a "vagueness dictionary" of words considered to indicate that the speaker is not certain about the material in his presentation. Examples of such words are: almost, generally, may be, many, and most. We hypothesized

that teachers with a high proportion of vagueness words would be ineffective lectures, that is, that there would be a negative correlation between the proportion of vagueness words and lecturing effectiveness. The following is an excerpt taken from a lecture with a high vagueness count:

> "...and the young author's name, although this is not too important thing to remember is that it was a young author who wrote this. I will put his name up on the board anyway. It is really not very important at all. MIHAJOV - that is the way you pronounce that word, Uh Mihajlov wrote those articles. And someone, he has done something that is fine someone very similar had done and there was another author whose name uh, uh, let us just remember there is another author. That one has spelling problems too. Two authors, two authors. One we know is Mihajlov, the other one wrote earlier in nineteen sixty-two. Both of them complained about conditions, especially in Russia. And this one was in prison because he wrote a book about conversations with Stalin and, I do not know if you have ever heard of the book. And this one also just recently has also been in prison."

Hiller also developed an "adherence-to-detail" dictionary consisting of all proper nouns and place designations in the original articles, and a "problem-issue" dictionary consisting of words such as "conflict," "divergent," and "issue," which might be used by teachers to highlight certain problems and issues.

In addition, a dictionary was made of "explaining links" as defined by Rosenshine, although our dictionary is not identical to the one developed by Rosenshine because we included words, regardless of their context, that might serve as explaining links. For example, in our dictionary we included all instances of words such as "to" and "since" as well as such words as "therefore" and "because." Although the coders in Rosenshine's study counted all instances of such words as "therefore" and "because," they relied upon context to decide which instances of words such as "to" and "since" should be counted.

The total list of words used in all the dictionaries was approximately 7,000. But many of these words appeared in more than one dictionary.

Results

To test the usefulness of computer analysis, a random sample of 15 was drawn from the group of 35 lectures on Yugoslavia, and a second sample of 15 on Thailand was drawn at random from the lectures of teachers not represented in the Yugoslavia sample. In other words, two groups of 15 lectures each were drawn, with no teacher represented in both groups.

The computer referred to the dictionaries to count the frequency and proportion of words in each category, and then correlated this proportion with the teacher's effectiveness score. Some of the correlations of interest are presented in Table 12. (For N = 15, an r of .514 is significant at the .05 level.)

The categories presented in Table 12 are those that correlated at least .30 in the same direction with effectiveness in explaining for both the Yugoslavia and the Thailand protocols.

The first nine categories listed in Table 12 came from thé Harvard Third Psychosociological Dictionary. In Categories 6, Medical, and 8, Sex Theme, the mean scores on these items were so low that undue chance factors may have been involved; no plausible explanation relating sex themes or medical terms to the effectiveness criterion suggested itself. The results suggested the speculations that successful explanations are communications involving task orientation for the pupils (Category 5, Academic), place orientation (Categories 1, Space Reference, and 2, Social Place), and relationships between elements (Categories 3, Avoid; 4, Get; 7, Sign Reject; and 9, Danger Theme). In general, the correlations of these categories were low enough to make such speculation dubious.

The results in Table 12 also bear upon the promise of the categories in the semantic differential approach of Holsti's Stanford Political Dictionary. In general, these categories had little value in accounting for variance in effectiveness in explaining.

Categories 14, Explaining Links 2, and 15, Explaining Links Total, are related to the "explaining links" described above. Here, as already noted, the measures of this variable were reduced to a count of words without regard to context. Consequently, although the resulting correlations remain high (rs = about .4), they have dropped below the level of significance found in Study III.

The results for Category 12, Problem-Issue, indicated that the more effective teachers used a higher proportion of words such as "conflict," "divergent," and "issue." Such words may have served to arouse attending behavior, or to focus the pupils' attention upon critical points.

Table 12

Frequencies of Words in Various Categories: Their Mean Frequencies and Correlations with Effectiveness in Explaining

Item No.	Category	Source and Definition	Word Sample	Yugoslavia (N = 15)		Thailand (N = 15)	
				Correlation	Mean Frequency	Correlation	Mean Frequency
1	Space Reference	Harvard Psy. III Reference to spatial dimensions	about, ahead, back	.509	75.9	.337	67.7
2	Social Place	Harvard Psy. III Political, social, economic locations	America, bedroom, cabin	.371	27.4	.350	29.0
3	Avoid	Harvard Psy. III Movement away from	abandon, absent,	.405	8.7	.541	11.5
4	Get	Harvard Psy. III Achieving action	afford, attain, beg	.454	16.7	.316	11.6
5	Academic	Harvard Psy. III	assignment, correct, teach	.394	30.6	.324	26.1
6	Medical	Harvard Psy. III	therapy, treatment, injury	-.560	1.1	-.347	.2
7	Sign Reject	Harvard Psy. III Words implying rejection	anger, betray, sulk	.319	33.3	.373	29.7
8	Sex Theme	Harvard Psy. III Reference to sex act	engagement, attentive, embrace	.749	4.5	.368	4.5
9	Danger Theme	Harvard Psy. III Connoting concern for danger	blast, warn, deviant	.342	9.4	.524	4.9

Table 12 (continued)

Item No.	Category	Source and Definition	Word Sample	Yugoslavia (N = 15)		Thailand (N = 15)	
				Correlation	Mean Frequency	Correlation	Mean Frequency
10	Affective -2 (proportional)	Holsti Words with middle rating on affective negative scale	complain, embarrass, slander	-.336	26.1	-.368	23.5
11	Strong	Holsti Weighted	competent, shatter, leader	-.032	393.1	.230	387.9
12	Problem-Issue	Words used by teachers in presenting issues	conflict, divergent, issue	.290	9.7	.551	11.0
13	Vagueness (proportional)	Indicating lack of precision	almost, generally, many	-.375	42.6	.249	49.9
14	Explaining Links 2	Rosenshine (Study III above	Contains only the words: if, then, to	.407	55.3	.407	56.2
15	Explaining Links Total	Rosenshine (Study III above)	All of Explaining Links 2 above, plus 15 other words, such as: consequently, because, since	.384	119.4	.372	112.5
16	Adherence to detail	All proper nouns and place designations in original article	Thailand, Tito, Hitler	.321	121.4	.476	111.6

The results for Category 16, Adherence to Detail, indicated that the more effective teachers in both groups used a greater proportion of proper nouns and place designations. This finding is difficult to interpret without further detailed study of the original transcripts. One possibility is that the less effective teachers used a greater number of pronouns in place of the proper nouns, and that such pronoun references detracted from the clarity achieved by using proper nouns.

Further Research on the "Vagueness Dictionary"

The initial research in the use of the computer to count instances of vagueness was expanded in a subsequent study by Hiller, Fisher, and Kaess (1968). In this investigation, Fisher developed a new computer program, SCORTXT, which is capable of counting instances of selected phrases in addition to single words. As a result of this new computer capability, a new vagueness dictionary was developed, consisting of 233 entries in several subdivisions. The subdivisions and examples of the new vagueness words and phrases are presented in Table 13. The validity of this new dictionary was tested on 32 lectures on Yugoslavia and 23 lectures on Thailand. In this study, the correlation between the proportion of vagueness words and phrases and the effectiveness-in-explaining criterion was -.59 for the Yugoslavia groups, and -.48 for the Thailand group; both *rs* are significant beyond the .02 level.

Discussion

This research has demonstrated that computer techniques can be developed to count certain aspects of classroom discourse. The development of computer programs which can count phrases appears to be a significant step beyond the first programs, which were limited to counting specific words. Such a new program has led to the discovery that the proportion of certain words and phrases classified as indicating "vagueness" is a signficant negative correlate of effectiveness in explaining. This initial finding appears promising and warrants future research. Particular subcategories of the vagueness words should be validated to determine which of the nine subcategories should be retained or dropped in future research. In addition, the computer can be used to validate the words and phrases that had been included in the subcategories on an a priori basis.

Table 13

Illustrated Vagueness Categories and Statistics

Category	Example	Number of Items	Mean Number Occurring
Ambiguous designation	all of this and things somewhere other people	39	4.7
Negated intensifiers	not all not many not very	48	1.2
Approximation	about as almost pretty much	25	2.3
"Bluffing" and recovery	a long story short anyway as you all know of course	27	8.3
Error admission	excuse me not sure maybe I made an error	14	1.3
Indeterminate quantification	a bunch a couple few some	18	10.3
Multiplicity	aspects factors sorts kinds	26	7.8
Possibility	may might chances are could be	17	8.0
Probability	probably sometimes ordinarily often frequently	19	2.0
	Totals	233	45.9

In short, much replicational and cross-validational work remains to be done on the categories of words and phrases studied thus far. Future work on these and other dictionaries that might show promise will be necessary to strengthen or discredit their significance as predictors of teacher effectiveness in explaining. The major limitation of the computer as an aid to analysis lies in the inability of the computer to determine context. Thus, although a human rater can distinguish between the use of the word "since" to indicate "because" and its use to indicate "after," a computer is unable to perform this task at present. Such a limitation may be only temporary, however, and may be overcome by imaginative and resourceful investigators.

References

Allen, W. H. Research on film use: Class preparation. A-V Comm. Rev., 1955, 3, 183-96.

Ausubel, D. P. The psychology of meaningful verbal learning. New York: Grune & Stratton, 1963.

Bellack, A. A., Kliebard, H. N., Hyman, R. T., & Smith, F. L. The language of the classroom. New York: Teachers College Press, 1966.

Brinton, J. E., & Danielson, W. A. A factor analysis of language elements affecting readability. Journalism Q, 1958, 35, 420-26.

Chall, J. S. Readability: An appraisal of research and application. Bureau of Educational Research Monographs, No. 54. Ohio State University, 1958.

Dale, E. Dale list of 3,000 familiar words. Bureau of Educational Research. Columbus, Ohio: Ohio State University, Educational Research Bulletin No. 27, 1948, 45-54.

Fortune, J. C., Gage, N. L., & Shutes, R. E. Generality of the ability to explain. Paper presented at the meeting of the American Educational Research Association, Chicago, February 1966.

Flanders, N. A. Teacher influence, pupil attitudes, and achievement. Coop. Res. Monograph No. 12, OE-25040. Washington, D. C.: U. S. Department of Health, Education, and Welfare, 1965.

Gage, N. L. Paradigms for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 94-141.

Getzels, J. W., & Jackson, P. W. The teacher's personality and characteristics. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 506-82.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954. Second edition.

Hiller, J. E. An experimental version of the effects of conceptual vagueness on speaking behavior. Unpublished doctoral dissertation, University of Connecticut, 1968.

Hiller, J. E., Fisher, G., & Kaess, W. A computer investigation of characteristics of teacher lecturing behavior. Paper presented at the meeting of the American Educational Research Association, Chicago, February 1968.

Holsti, O. R. Content analysis research in the social sciences. Paper presented at the IBM-Texas A&M Conference on Computers in Humanistic Research, College Station, Texas, November 1966.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. Experiments on mass communication. Princeton, N. J.: Princeton University Press, 1949.

Klare, G. R. The measurement of readability. Ames, Iowa: Iowa State University Press, 1963.

Lumsdaine, A. A. Instruments and media of instruction. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 583-682.

May, M. A., & Lumsdaine, A. A. Learning from films. New Haven, Conn.: Yale University Press, 1963.

Medley, D. M., & Mitzel, H. E. Some behavioral correlates of teacher effectiveness. J. Educ. Psychol., 1959, 50, 239-46.

Meux, M., & Smith, B. O. Logical dimensions of teaching behavior. In B. J. Biddle & W. J. Ellena (Eds.), Contemporary research on teacher effectiveness. New York: Holt, Rinehart, & Winston, 1964. Pp. 127-64.

Mosteller, F., & Wallace, D. L. Inference and disrupted authorship: The federalist. Reading, Mass.: Addison-Wesley, 1964.

Nuthall, G. A., & Lawrence, P. J. Thinking in the classroom: The development of a method of analysis. Wellington, New Zealand: New Zealand Council for Educational Research, 1965.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. The measurement of meaning. Urbana, Ill.: University of Illinois Press, 1957.

Page, E. B. Grading essays by computer. Phi Delta Kappan, 1966, 47, 238-43.

Petrie, C. R., Jr. Informative speaking: A summary and bibliography of related research. Speech Mon., 1963, 30, 79-92.

Rosenshine, B. Objectively measured behavioral predictors of effectiveness in explaining. Unpublished doctoral dissertation, Stanford University, 1968.

Rothkopf, E. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. Amer. Educ. Res. J., 1966, 3, 241-49.

Smith, B. O., & Meux, M. A study of the logic of teaching. U. S. Department of Health, Education, and Welfare, Coop. Res. Prog. Proj. No. 258 (7257). Urbana, Ill.: Bureau of Educational Research, College of Education, University of Illinois, n.d.

Soar, R. S. An integrative approach to classroom learning. Final Report, Public Health Service Grant No. 5-R11 MH 01096 and National Institute of Mental Health Grant No. 7-R11-MH 02045. Philadelphia, Pa.: Temple University, 1966.

Spaulding, R. L. Achievement, creativity, and self-concept correlates of teacher-pupil transactions in elementary schools. Urbana, Ill.: University of Illinois, 1963. (Coop. Res. Proj. No. 1352, U. S. Office of Education.)

Stolurow, L. M., & Newman, J. R. A factorial analysis of objective features of printed language presumably related to reading difficulty. J. Educ. Res., 1959, 52, 243-51.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. The general inquirer: A computer approach to content analysis. Cambridge, Mass.: Massachusetts Institute of Technology Press, 1966.

Swift, L. F. Explanation. In B. O. Smith & R. H. Ennis (Eds.), Language and concepts in education. Chicago: Rand McNally, 1961. Pp. 179-94.

Thyne, J. M. The psychology of learning and techniques of teaching. London: University of London Press, 1963.

Travers, R. W. W., et al. Research and theory related to audiovisual information transmission. Interim Report, U. S. Department of Health, Education, and Welfare, Office of Education Contract No. 3-20-003. Salt Lake City, Utah: University of Utah, Bureau of Educational Research, 1964.

Unruh, W. R. The modality and validity of cues to lecture effectiveness. Unpublished doctoral dissertation, Stanford University, 1967.